An Analysis of Earth Science Data Analytics Use Cases

AGU/15
IN23C
-1736

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

Chung-Lin Shie[1,2], Steve Kempler[1]

[1]NASA Goddard Earth Science Data and Information Services Center (GES DISC)
[2]University of Maryland, Baltimore County
chung-lin.shie-1@nasa.gov

Analyzing how Earth Science Data Analytics (ESDA) are used reveals foundation for ESDA goals categorization

Abstract

***Earth Science Data Analytics** (ESDA) is the process of examining **large amounts of data of various types** to uncover **hidden patterns**, **unknown correlations** and other **useful information**. It can include **Data Preparation**, **Data Reduction**, and **Data Analysis**. Through work associated with the **Earth Science Information Partners (ESIP) Federation**, a collection of **use cases** have been analyzed for the purpose of extracting the **types of Earth science data analytics** employed, and **requirements for tools and techniques** yet to be implemented, based on **use case needs**. **ESIP generated use case template**, **ESDA use cases**, **use case types**, and preliminary **use case analysis** (this is a work in progress) will be presented.*

Earth Science Data Analytics Definition

The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

This encompasses:

- ***Data Preparation** – Preparing heterogeneous data so that they can be jointly analyzed*
- ***Data Reduction** – Correcting, ordering, and simplifying data in support of analytic objectives*
- ***Data Analysis** – Applying techniques/methods to derive results*

Use Cases Template

*a. **Use Case Title***
*b. **Author/Company/Email** - author of the use case*
*c. **Actors/Stakeholders/Project URL and their roles and responsibilities***
*d. **Use Case Goal** - What is the goal of the Earth science data analytics?*
*e. **Use Case Description** - This yields more details regarding how data analytics is utilized.*
*f. **Current technical issues/requirements** to take into account that may impact needed data analytics. These can include:*

- ***Data Source** (distributed/centralized)*
- ***Volume** (size)*
- ***Velocity** (e.g. real time)*
- ***Variety** - Bringing distributed heterogeneous data together*
- ***Veracity** (Robustness Issues) / Data Quality*
- ***Visualization***
- ***Computer (System), storage, networking***
- ***Specialized Software***
- ***Current Data Analytics tools** applied*

*g. **Data Analytics Challenges (Gaps)** - Identifying known data analytics challenges, roadblocks, areas needing attention*
*h. **Type of User** - Taken from the ESIP Use Analysis Study, types of user performing use case.*
*i. **Dominant Data Analytics Skills Needed** - Skills needed to perform use case analytics*
*j. **Science Research Areas** - NASA Earth science research areas (http://science.nasa.gov/earth-science/focus-areas/)*
*k. **Societal Benefit Areas** – GEOS or NASA Applications (http://appliedsciences.nasa.gov)*
*l. **Potential for and/or issues for generalizing this use case***
*m. **More Information and relevant URLs** (e.g. who to contact or where to go for more information)*

Goals of Earth Science Data Analytics

1. ***To calibrate data***
2. ***To validate data** (note it does not have to be via data intercomparison)*
3. ***To assess data quality***
4. ***To perform coarse data preparation** (e.g., subsetting, data mining, transformations, recover data)*
5. ***To intercompare data** (e.g., any data intercomparison; Could be used to better define validation/quality)*
6. ***To tease out information from data***
7. ***To glean knowledge from data and information***
8. ***To forecast/predict phenomena** (e.g., Special kind of conclusion)*
9. ***To derive conclusions** (e.g., that do not easily fall into another type)*
10. ***To derive new analytics tools***

Conclusions (thus far, with our limited number of use cases):

***I.** For **Earth Science**, defining results oriented Data Analytics types are more appropriate for categorizing Earth science data analytics…*
- *They **accommodate Earth science use cases which are typically results oriented***
- *They **invite better defined data analytics tools and techniques that address user goals***

II.** Most **ESDA use cases** tend to focus on **data intercomparison, deriving new products, forecasting/predicting, and deriving conclusions

***III.** Most use cases were not identified to glean knowledge from data/information. Perhaps some use cases were not recognized as such (more analysis needed)*

IV. Distributed data sources**, and **data heterogeneity** are **persistent characteristics

***V. Velocity** issues are not significant (thus far)*

***VI. ESDA challenges** provide interesting problems for data analytics tool/technique developers to ponder*

***VII.** If any, use case **5** and **16** provides the true **Big Data** problem*

	Earth Science Data Analytics Goals										*Other Significant Earth Science Data Analytics Considerations*							Current data analytics	Data Analytics
Use Cases	1	2	3	4	5	6	7	8	9	10	Data sources	Volume	Velocity	Variety	Veracity	Visualization	Specialized s/w	tools applied	Challenges
1 MERRA Analytics Services: Climate Analytics-as-a-Service										√	Distributed					For Mapping		Cloudera MapReduce	
2 MUSTANG QA: Ability to detect seismic instrumentation problems			√	√				√			Centralized	100's TB --> PB		Uniform	Problematic		scheduler, SQL	R, Matlab, Excel, PQLX	Large ds; erroneous data
3 Inter-calibrations among datasets	√	√			√														MIICII, XML
4 Inter-comparisons between multiple model or data products					√						Centralized	Huge		Heterogeneous		To Identify event			
5 Sampling Total Precipitable Water Vapor using AIRS and MERRA		√			√						Collocated			Heterogeneous		To detect differences		Sampling, Gridding	
6 Using Earth Observations to Understand and Predict Infectious Diseases								√	√		Distributed	Large		Heterogeneous		Data exploration, findings	db, math/stat modeling	Regression Modeling; Machine Training; Neural Network; R	Data heterogeneity; data/results validation
7 CREATE-IP - Collaborative REAnalysis Technical Environment - Intercomparison Project					√						Distributed	up to 1 PB		Different formats	Depends on input	WMS, UV-CAT, ArcGIS		Anomaly correction	Volume; Data heterogeneity
8 The GSSTF Project (MEaSUREs-2006)						√					Distributed	~1 TB		Heterogeneous	Depends on input		Intercomparison, collocation	Bulk flux formula, EOF	Large data inputs/outputs
9 Science- and Event-based Advanced Data Service Framework at GES DISC					√		√			√	Distributed			Diverse data		Show event		Giovanni, OpENDAP, Panoply	Virtual collection/Data list
10 Risk analysis for environmental issues								√			Distributed			Diverse data					Determine model output suitability
11 Aerosol Characterization					√				√		Distributed	Huge		Heterogeneous	Part of analysis	Customized	Developed as needed		Reliable pattern recognition
12 Creating One Great Precipitation Data Set From Many Good Ones						√					Distributed		Near real time	Diverse data	Can be a problem		Intercomparison; morphing	Kalman filtering technique	Intercalibrate datasets to produce best data
13 Reconstructing Sea Ice Extent from Early Nimbus Satellites	√			√							Single source	Large # of records			Very problematic				Unreadable tapes = not automated
*14 DOE-BER AmeriFlux and FLUXNET Networks **						√			√		Distributed			Diverse data		Graphs and 3D surfaces	EddyPro, python, Matlab, neural networks	Data mining, interpolation, fusion, R	Translation across diverse datasets
*15 DOE-BER Subsurface Biogeochemistry Scientific Focus Area **								√			Distributed			Diverse data	Very problematic	To understand data	PFLOtran, postgres, NEWT	Data mining, interpolation, fusion	Translation across diverse datasets
*16 Climate Studies using the Community Earth System Model at DOE's NERSC center **								√	√	√	Distributed	up to 30 PB	42 GByles/sec	Diverse data		To understand data	PIO, NCL, NCO, parallel NetCDF	Data reduction; analysis near archive	A true Big Data problem
*17 Radar Data Analysis for CReSIS **						√					Single source	~0.5 PB per year			Needs analysis		Matlab, MapReduce, MPI, GIS	Signal/Image processing	Immature image processing algorithms
*18 UAVSAR Data Processing, Data Product Delivery, and Data Service **						√					Centralized			2 main types		GIS	ROI_PAC, FGeoServer, GDAL		Human inspection needed

*** Borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php]**

Acknowledgements: Thanks to the work of the Earth Science Information Partners (ESIP) Federation, Earth Science Data Analytics Cluster